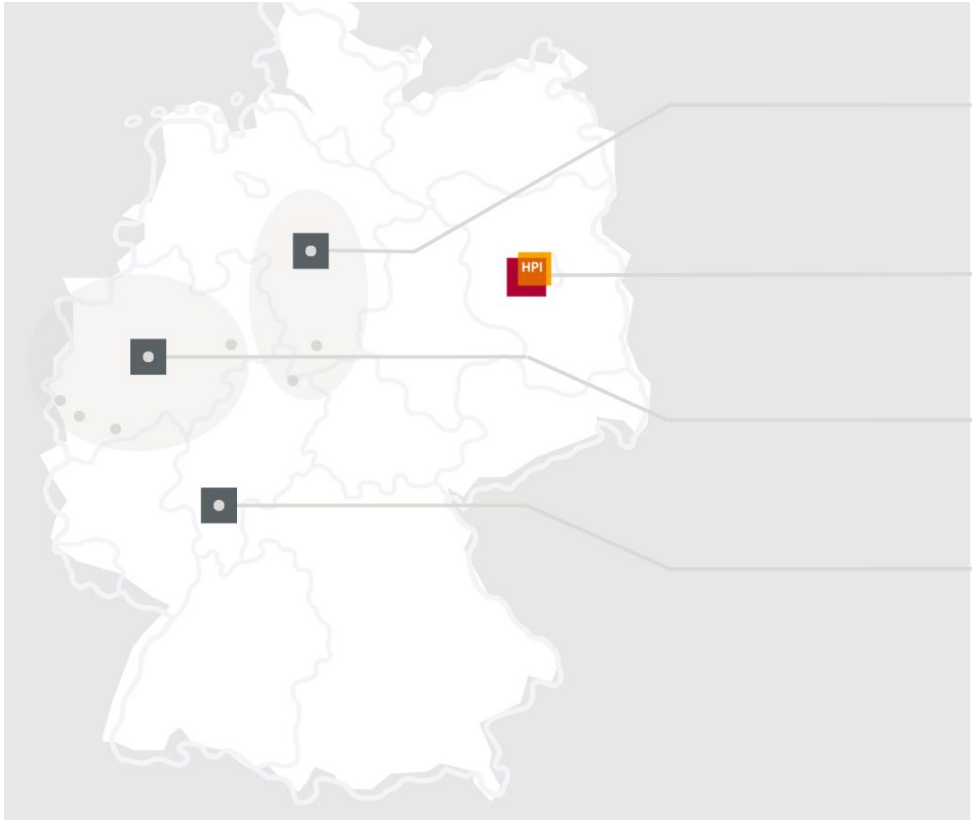
Image Dataset Curation
Workshop III

Antonio Rueda-Toicen
AI Engineer

KI Service Zentrum, Hasso Plattner Institute
February 2024

SPONSORED BY THE

https://hpi.de/kisz/home.html



KISSKI

Hannover I Göttingen I Kassel
Sensitive and critical infrastructures
Focus: Medicine & Energy

KISZ BB

Berlin I Brandenburg
Educational and advisory services
Use of AI in business and society

West AI

Dortmund I Bonn I Jülich I Aachen I Paderborn
Large and transferable AI models

hessian AI

Service Center

Darmstadt
Explainability, generalizability
and contextual adaptation

https://hpi.de/kisz/home.html

Agenda

- Brief intro and learning objectives
- Overview of parts I and II of this workshop series
  - Understanding image embeddings
  - Scraping images from Google Images
- Classifying images with pretrained Resnets
- Multi-label vs single label classification
- Fine-tuning a Resnet with FastAI
- Review questions and discussion

What we expect you to have

- Some Python knowledge
- Curiosity :)

Learning objectives

At the end of the first workshop you will be able to:

- Describe use cases for image similarity in dataset curation
- Scrape images from Google Images or Bing
- Generate embeddings for images using a pretrained neural network
- Compare image pairs using cosine similarity
- Visualize embeddings in 3D using Tensorboard

Learning objectives

At the end of the second workshop you will be able to:

- Visualize image neighborhoods with k-nn
- Cluster images using k-medoids
- Select representative images

Learning objectives

At the end of the third workshop you will be able to:

- Classify images with a pre-trained Resnet
- Fine-tune a Resnet for custom classes

How are we doing this workshop

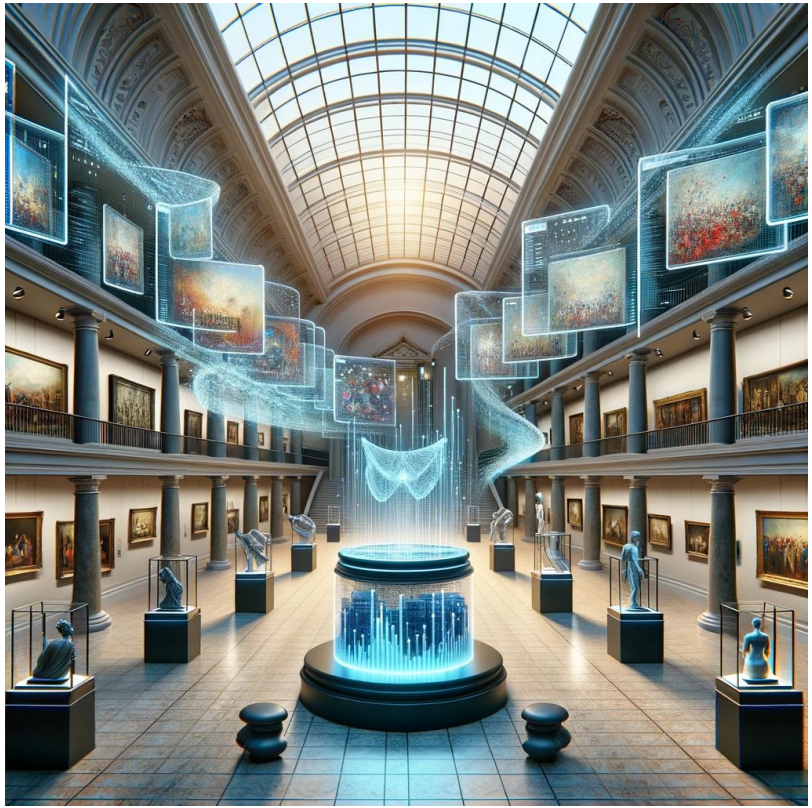
- We **type** most of the Python code from the tutorial notebooks in Google Colab

Repository:https://github.com/KISZ-BB/image-dataset-curation-workshops

What you need

- A Google user account
- A Google Drive account with enough free space
- Google Chrome or Firefox

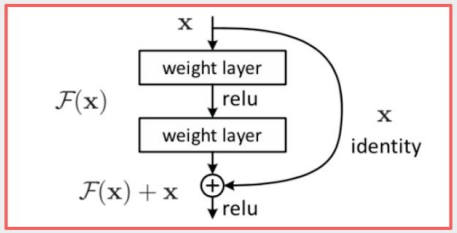
What is dataset curation?



- We want to make data **accurate** and **relevant**
- We clean, deduplicate, and label
- This is similar to what museum curators do

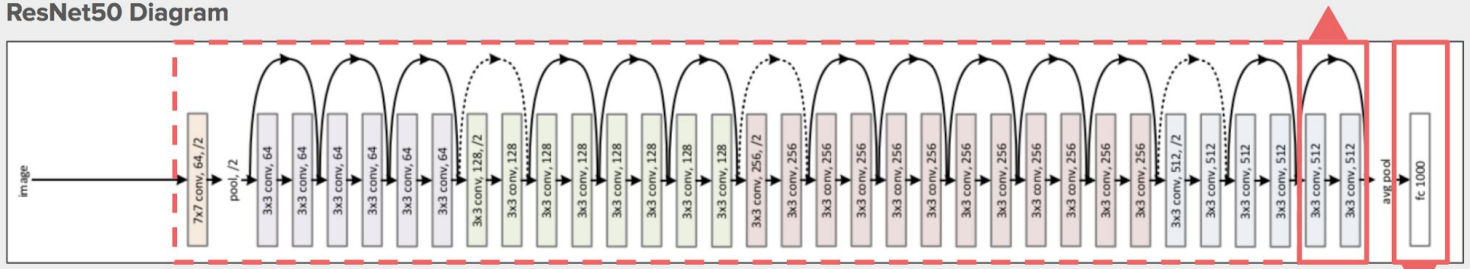
Resnet

Retrain ResNet50

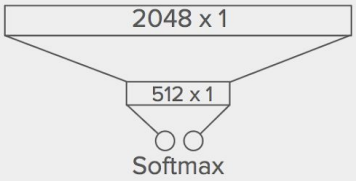


Residual Learning Block

ResNet50 Diagram



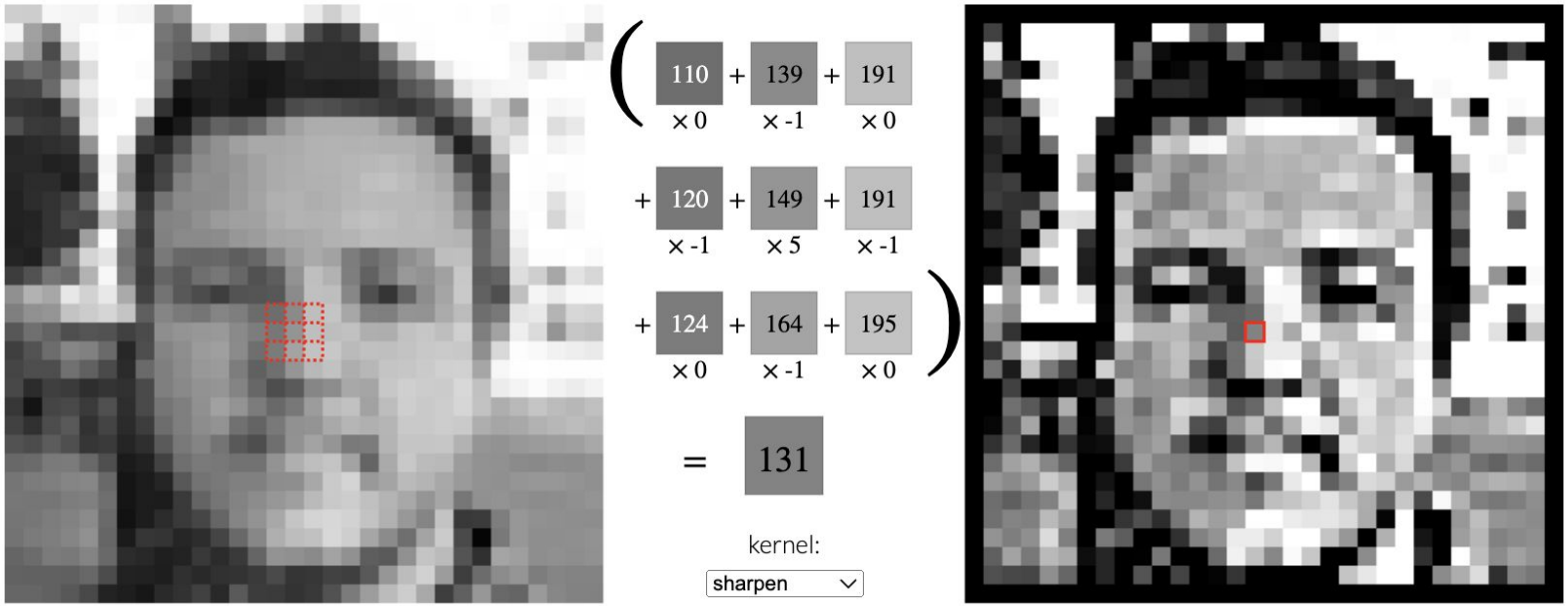
Re-architect fully-connected layers



The ImageNet dataset

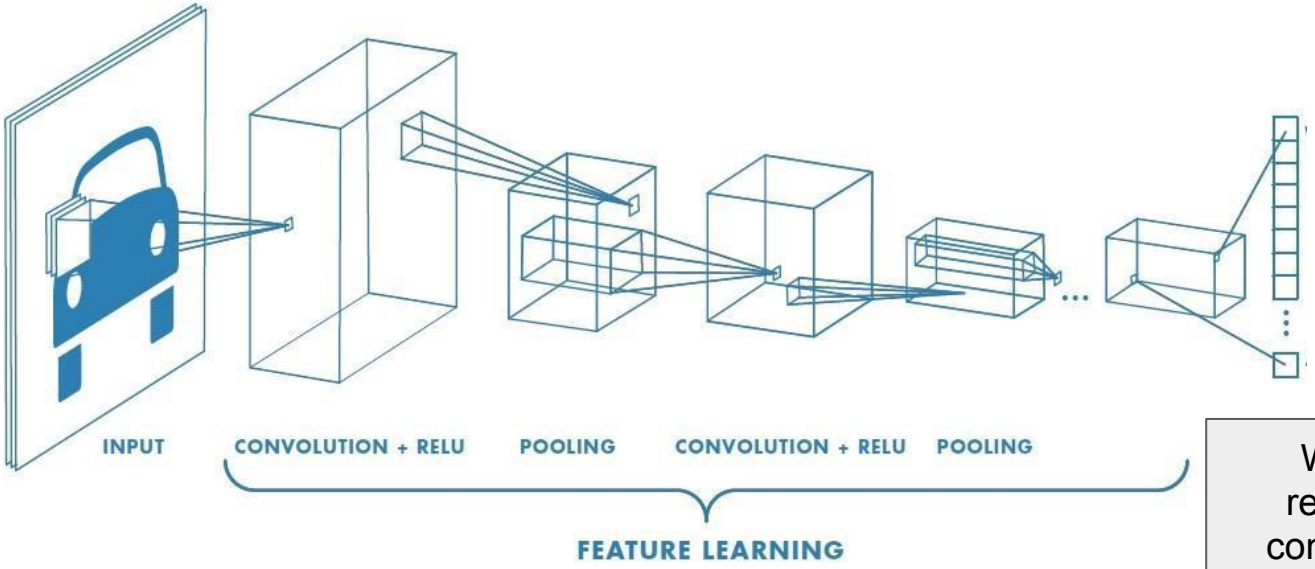
https://www.image-net.org/

Intuitions about convolutions



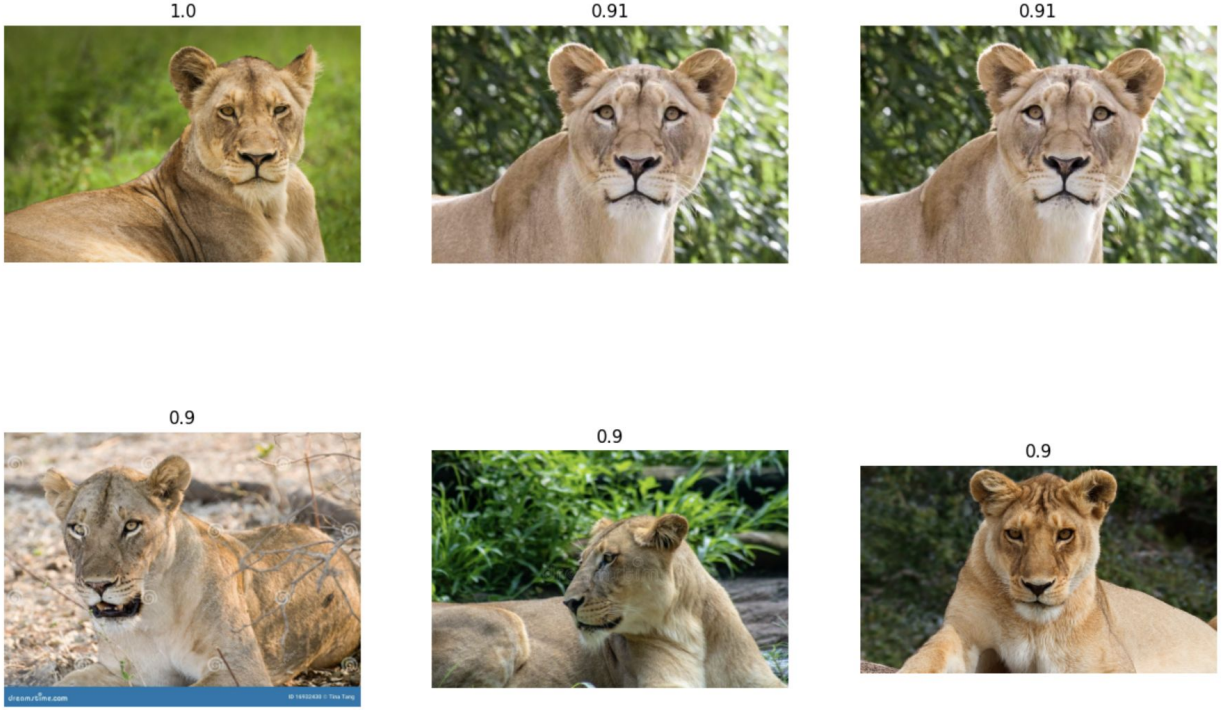
https://setosa.io/ev/image-kernels/

Image Embeddings from Convolutional Networks



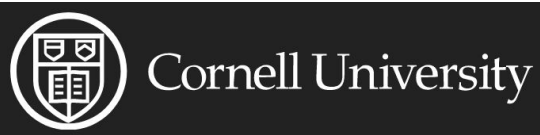
We use the flattened representation from all convolutions and pooling

Image neighborhoods



Explore the Colab notebook

What a pretrained network already knows



arXiv.org > cs > arXiv:1311.2901

Computer Science > Computer Vision and Pattern Recognition

[Submitted on 12 Nov 2013 (v1), last revised 28 Nov 2013 (this version, v3)]

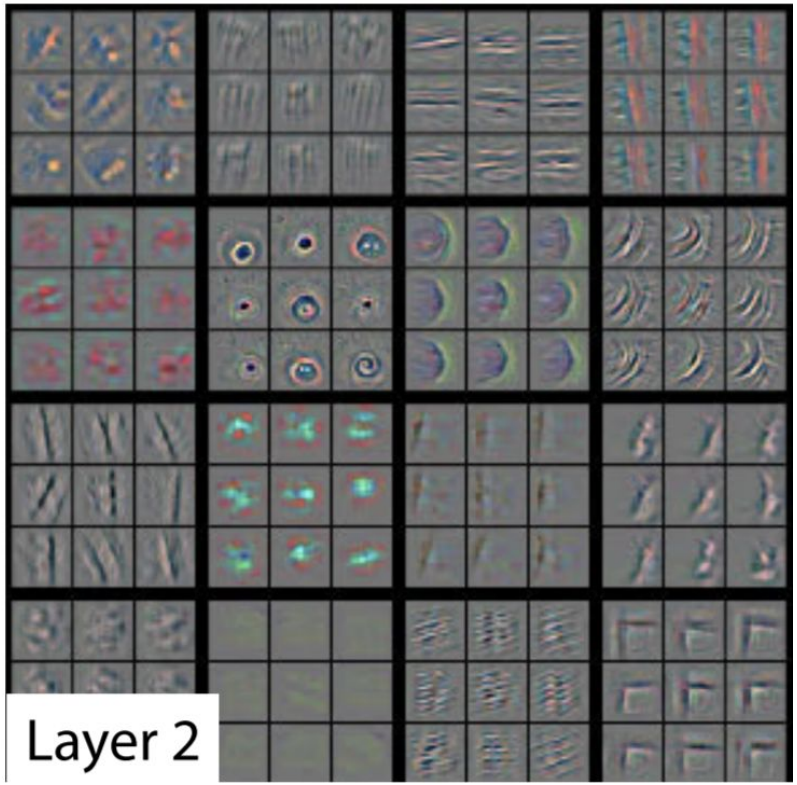
Visualizing and Understanding Convolutional Networks

Matthew D Zeiler, Rob Fergus

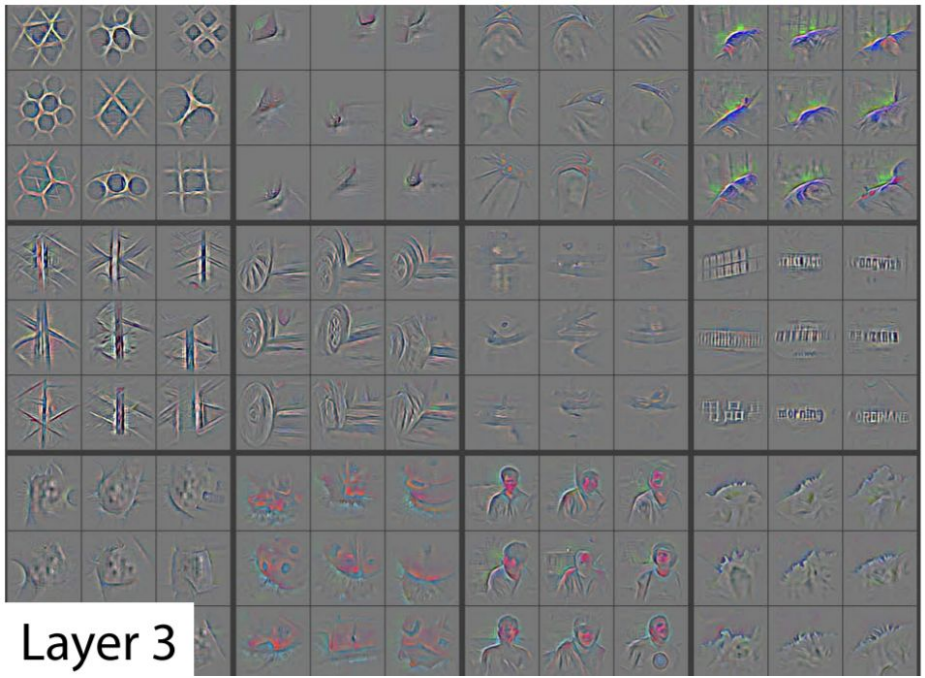
https://arxiv.org/pdf/1311.2901.pdf

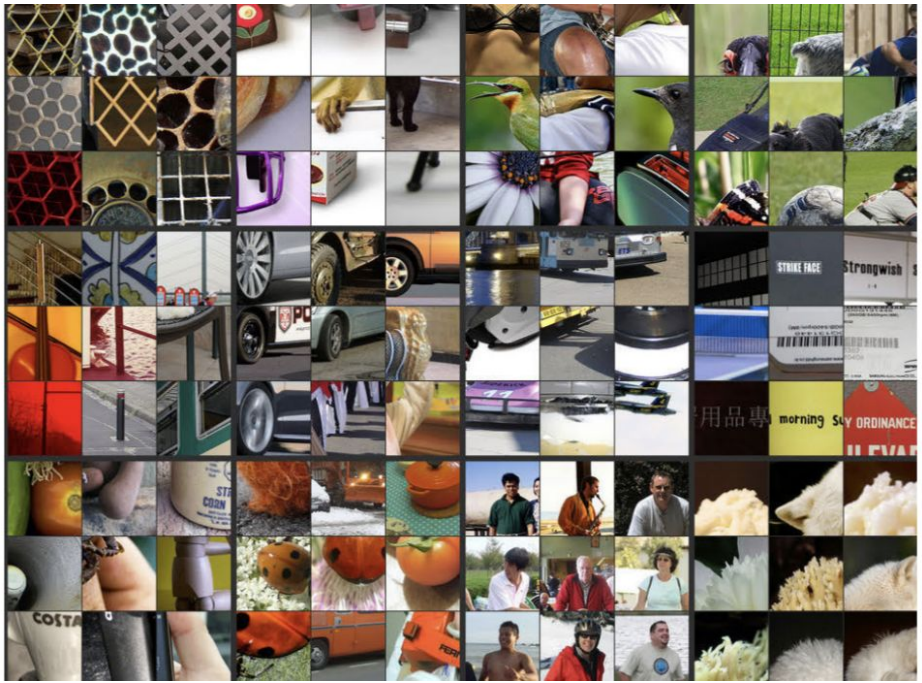
What a pretrained network already knows

Layer 1

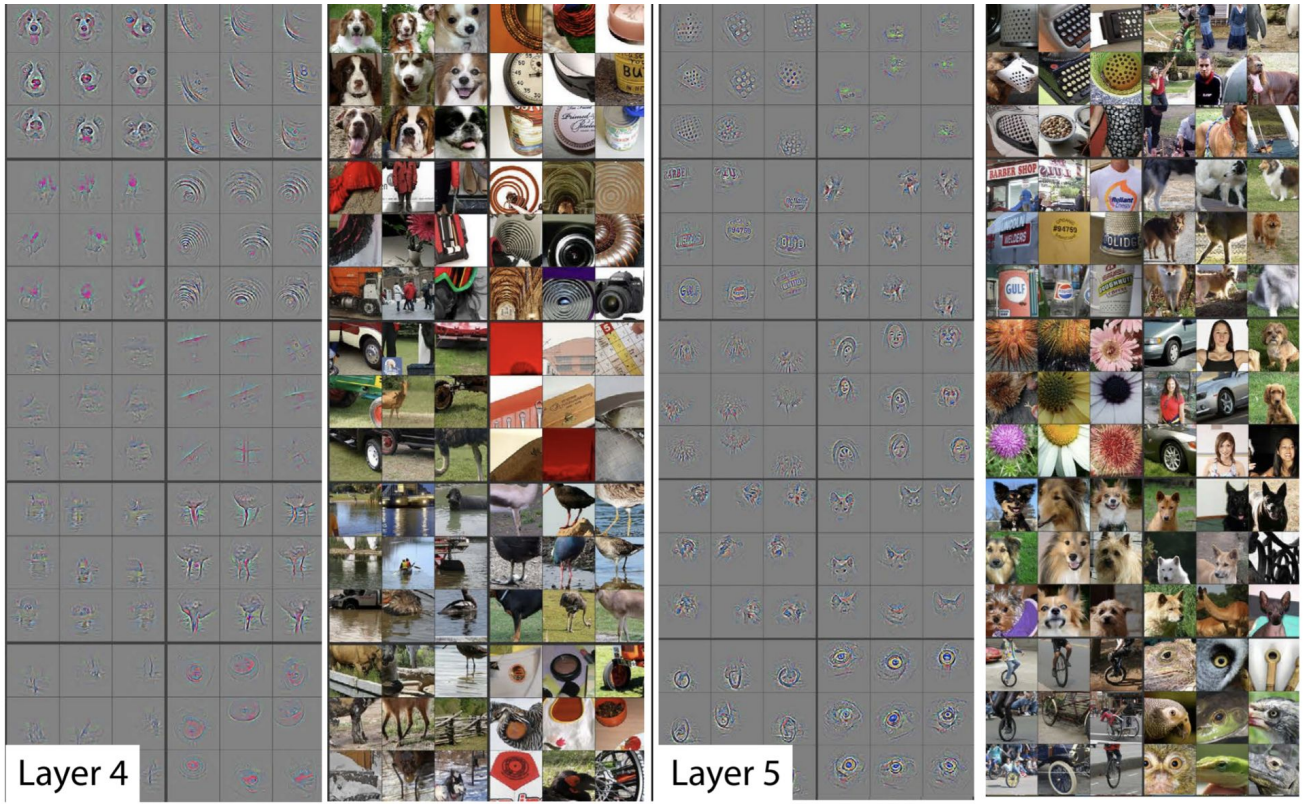


What a pretrained network already knows

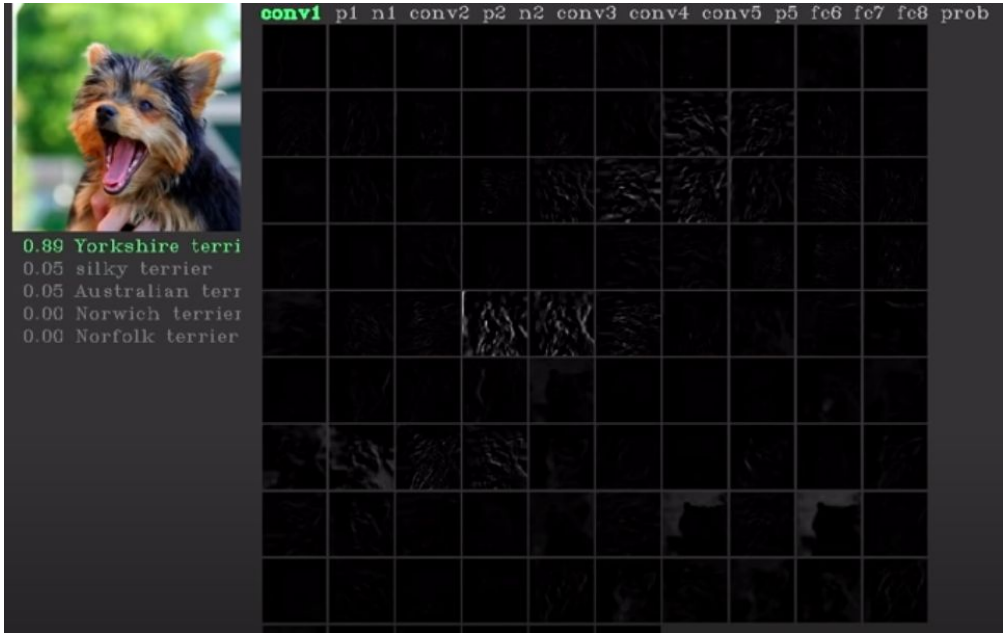




What a pretrained network already knows



What a pretrained network already knows



Visualizing the layers of a convolutional network trained on Imagenet

Lion or not a lion?

“If you torture the data long enough, it will confess to anything” Ronald Coase, economist

- Was it a lion?

Class Activation Mapping (CAM)

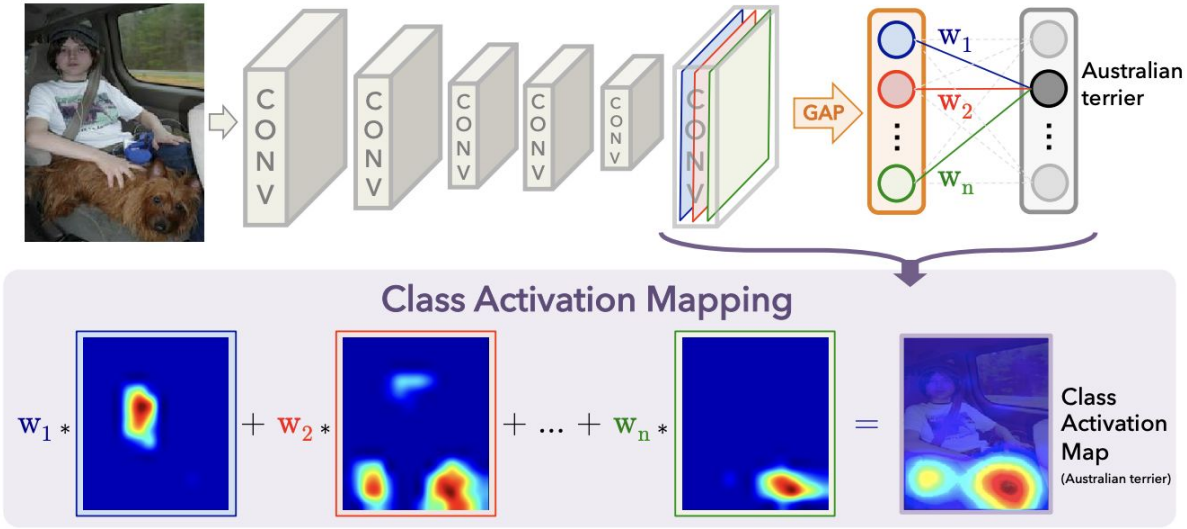
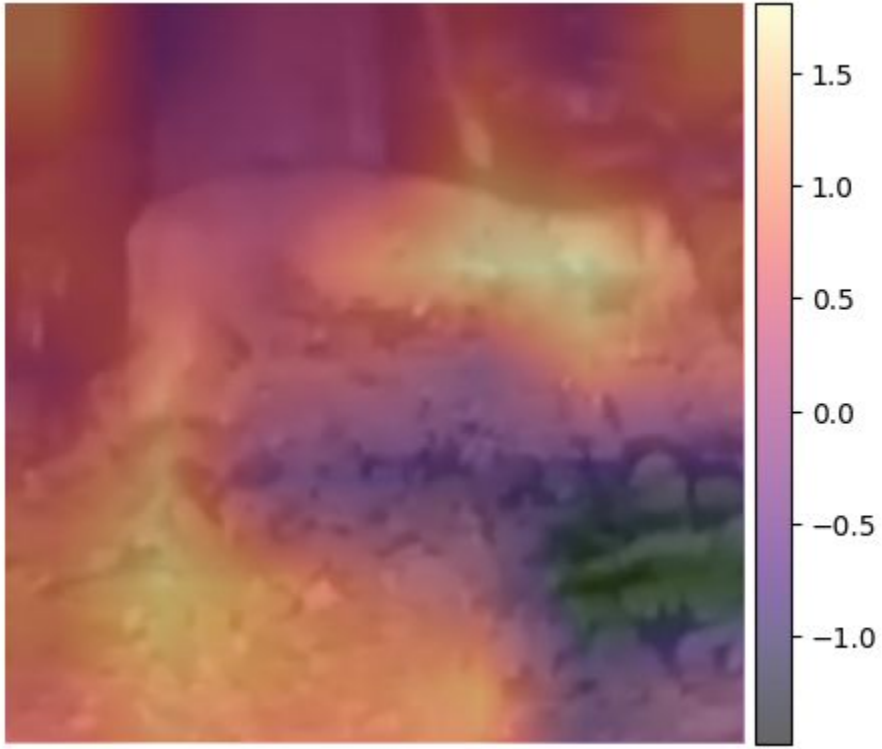


Figure 2. Class Activation Mapping: the predicted class score is mapped back to the previous convolutional layer to generate the class activation maps (CAMs). The CAM highlights the class-specific discriminative regions.

https://arxiv.org/pdf/1512.04150.pdf

Spurious correlations come from the training data

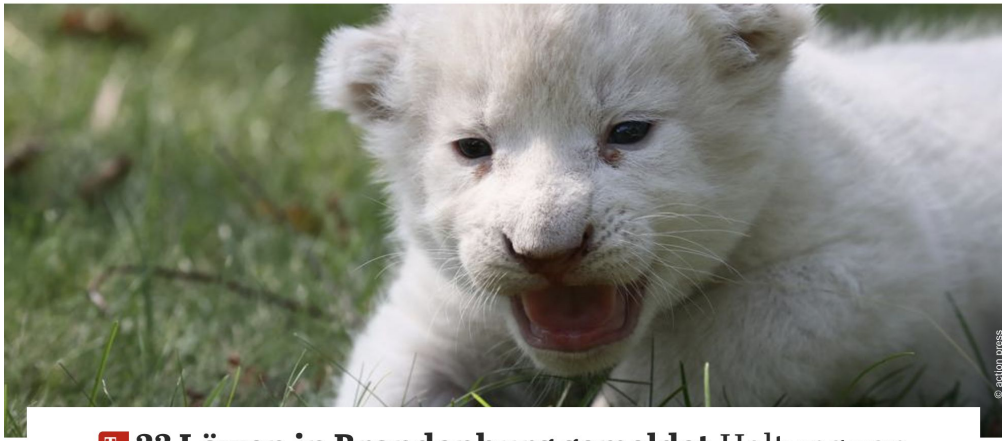


Here the model associates the **front of the animal** and the **vegetation** with the classes **'boar' (p=0.53)** and '**male lion**' **(p=0.66)**

Explore the notebook

- Which classes do you get?

- How does the CAM map look like?



T+ 23 Löwen in Brandenburg gemeldet Haltung von Großkatzen – das ist die Rechtslage in Berlin und der Region

Anders als in Brandenburg ist die private Wildtierhaltung in Berlin – mit einigen Ausnahmen – verboten. Wer es trotzdem tut, muss mit einer hohen Strafe rechnen.

Von Alexander Fröhlich und Daniel Böldt
21.07.2023, 11:43 Uhr

23 privately owned lions in Brandenburg!

Review questions

- How many classes are in the ImageNet dataset?
- What is transfer learning?
- Which part of the network is most often changed for fine-tuning?
- Which final activation function should be used for single label multiclass classification?
- Which final activation function should be used for multilabel multiclass classification?
- Can you explain the differences between train, validation, and test sets? What are the issues of having duplicates between these sets?
- Why are the performance metrics obtained on the dirty dataset **useless**?
- *Was it a lion?* 🤔

Join us for the next workshops!

- AI Service Center - Berlin Brandenburg

Next topics:

- **Zero-shot classification with CLIP**
- **Using classifiers to clean a dataset (Cleanlab)**
- **Meta's Segment Anything**
- **Segmentation and object detection with Detectron2, YOLOv8**
- **Using Qdrant as a vector database for images and text**
- **Deployment of a computer vision system with FastAPI and Docker**